

Threshold Pivoting for Dense LU Factorization

Neil Lindquist, Mark Gates, Piotr Luszczek, Jack Dongarra

ScalAH Workshop, 2022

Pivoting in Dense LU

- Needed for accuracy
 - Partial row pivoting used in practice
- Can add significant overhead
 1. Adds extra synchronizations
 2. Requires moving data to exchange rows

Pivoting constraints

- Partial Pivoting

$$|a_{ii}| \geq |a_{ji}| \quad i \leq j \leq n$$

Pivoting constraints

- Partial Pivoting

$$|a_{ii}| \geq |a_{ji}| \quad i \leq j \leq n$$

- Threshold Pivoting

$$|a_{ii}| \geq \tau |a_{ji}| \quad i \leq j \leq n$$
$$0 \leq \tau \leq 1$$

Accuracy

- Growth factor is main term in backward error bound
 - Growth in factorization \Rightarrow cancellation error

Accuracy

- Growth factor is main term in backward error bound
 - Growth in factorization \Rightarrow cancellation error
- Worst case: exponential growth

$$\rho \leq (1 + \tau^{-1})^{n-1}$$

Accuracy

- Growth factor is main term in backward error bound
 - Growth in factorization \Rightarrow cancellation error
- Worst case: exponential growth
$$\rho \leq (1 + \tau^{-1})^{n-1}$$
- Average case: ?

Accuracy

- Growth factor is main term in backward error bound
 - Growth in factorization \Rightarrow cancellation error
- Worst case: exponential growth
$$\rho \leq (1 + \tau^{-1})^{n-1}$$
- Average case: ?
- Growth of threshold pivoting given growth of partial pivoting

Growth: partial vs threshold pivoting

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 1 \\ -1 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & -1 & \cdots & 1 & 1 \\ -1 - \delta & -1 & \cdots & -1 & 1 \end{bmatrix}$$

$$0 < \delta < \min(\tau^{-1} - 1, 1)$$

Partial: $\rho \approx 2$

Threshold: $\rho \approx 2^{n-1}$

Growth: partial vs threshold pivoting

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 1 \\ -1 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & -1 & \cdots & 1 & 1 \\ -1 - \delta & -1 & \cdots & -1 & 1 \end{bmatrix}$$

$$0 < \delta < \min(\tau^{-1} - 1, 1)$$

Partial: $\rho \approx 2$

Threshold: $\rho \approx 2^{n-1}$

$$\begin{bmatrix} -1 & -1 & \cdots & -1 & 1 \\ -1 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & -1 & \cdots & 1 & 1 \\ 1 + \delta & 0 & \cdots & 0 & 1 \end{bmatrix}$$

$$0 < \delta < \min(\tau^{-1} - 1, 1)$$

Partial: $\rho \approx 2^{n-1}$

Threshold: $\rho \approx 2$

Avoiding inter-process communication

- Distributed codes → low network bandwidth

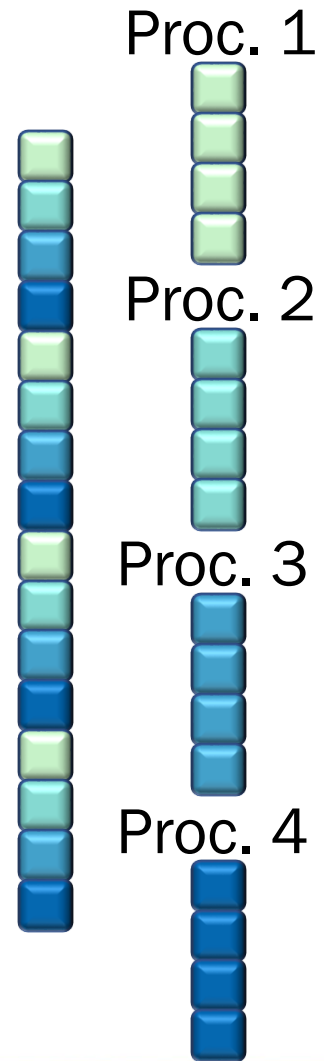
Avoiding inter-process communication

- Distributed codes \rightarrow low network bandwidth
- Assume: a_{ii}, a_{ji} on same process
iff a_{ik}, a_{jk} on same process $\forall k$
 - E.g., 2d block-cyclic

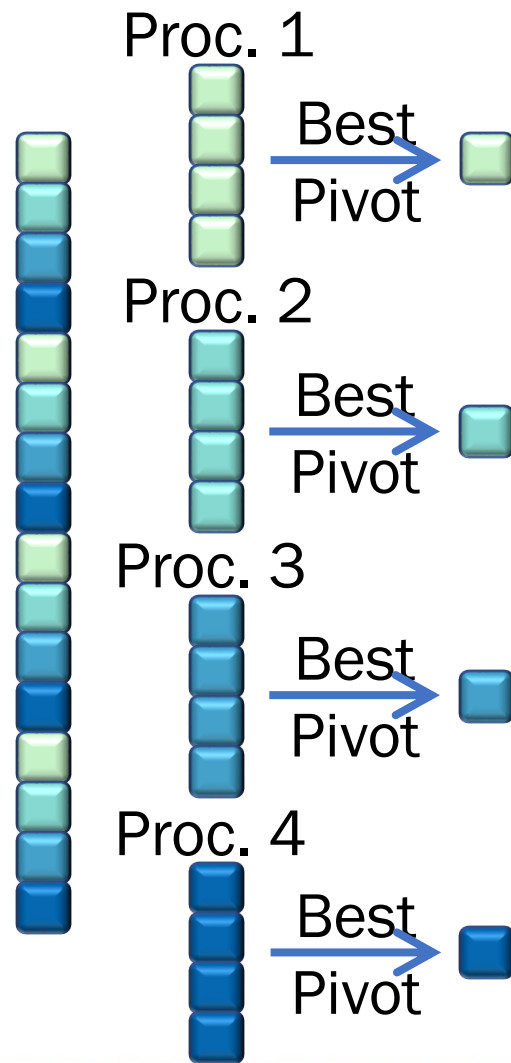
Avoiding inter-process communication



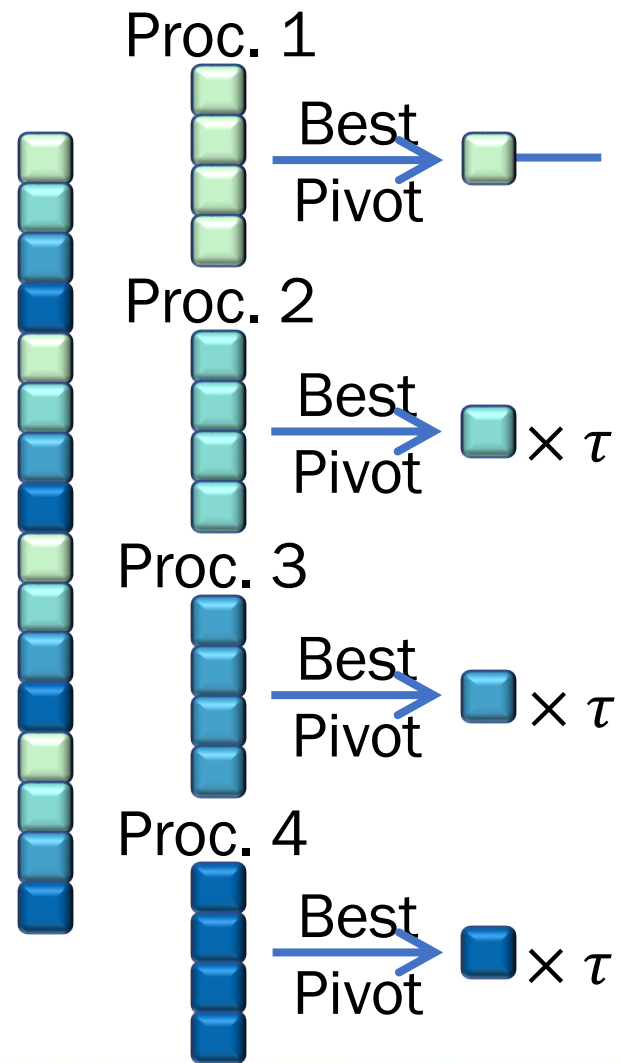
Avoiding inter-process communication



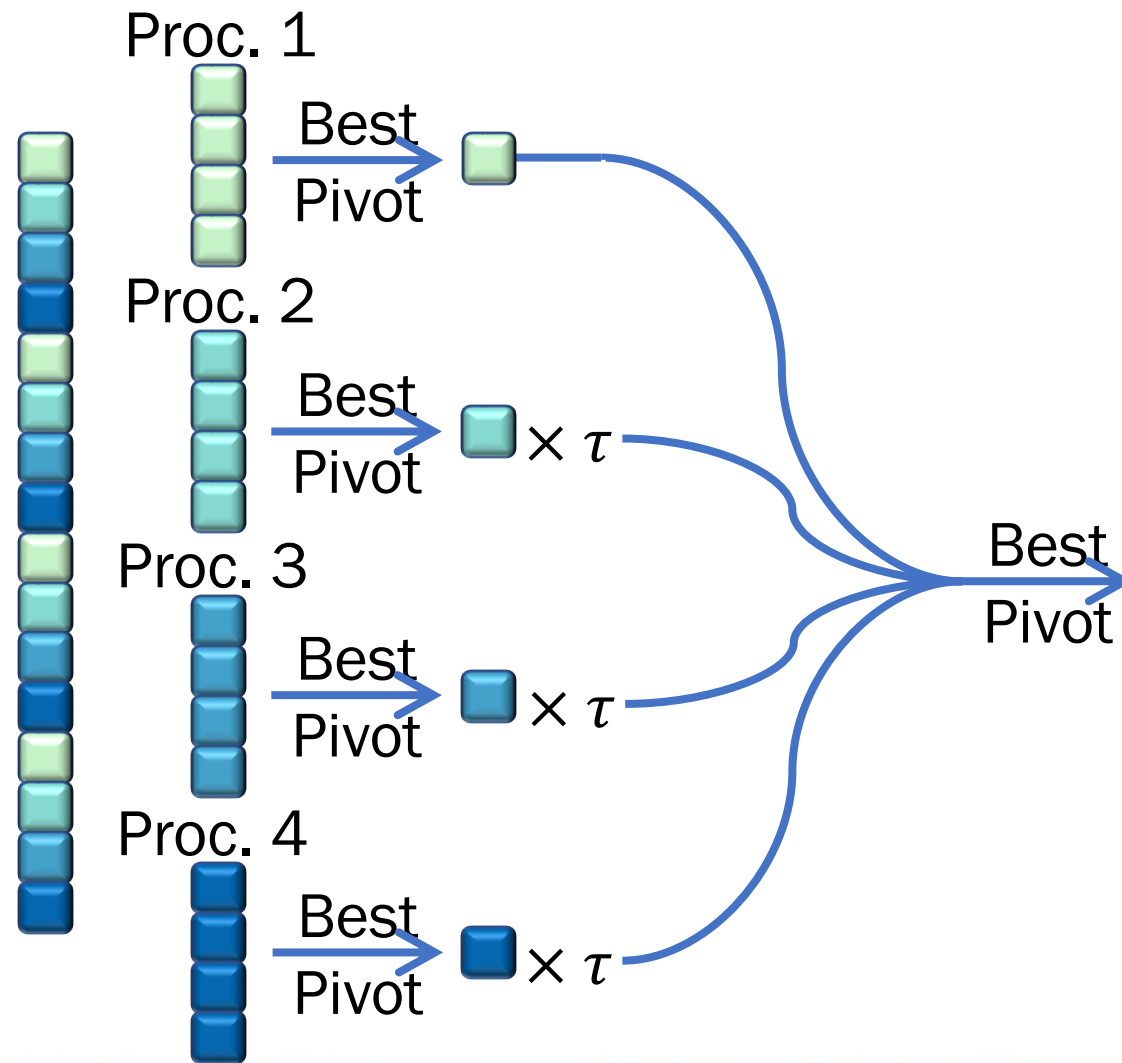
Avoiding inter-process communication



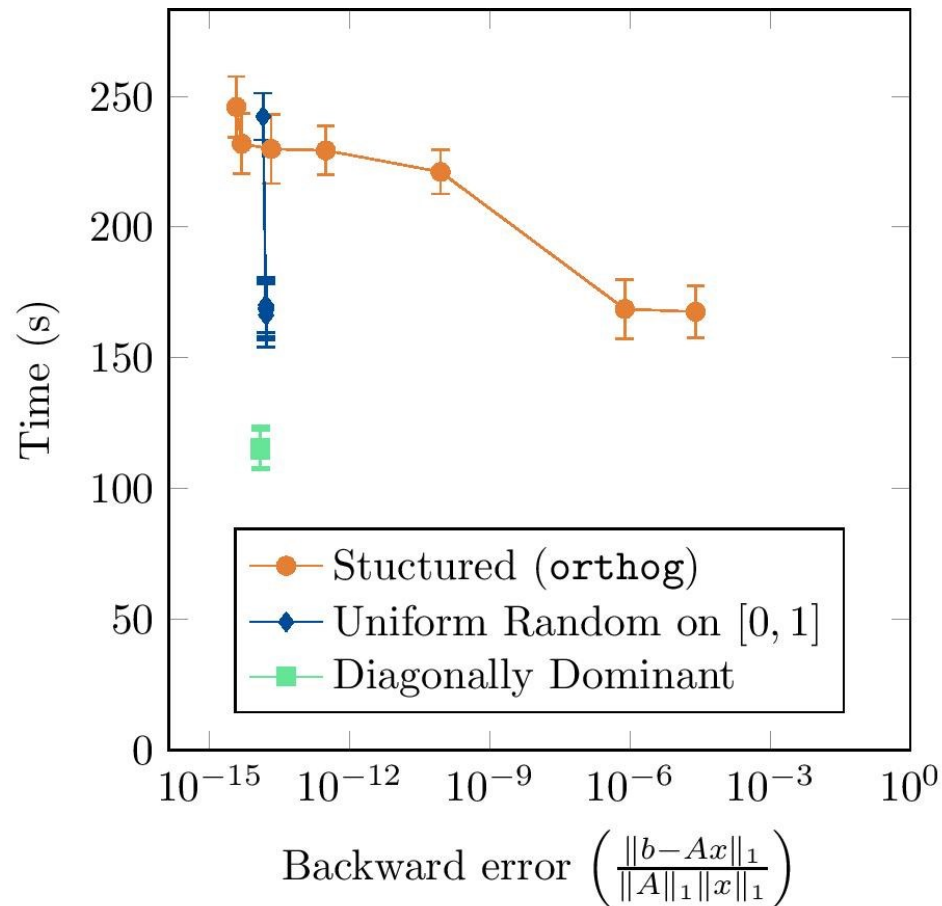
Avoiding inter-process communication



Avoiding inter-process communication



Effect on performance



- 8 nodes of Summit
- SLATE w/ target=device
- $n = 225\,000$; nrhs = 10
- Double precision
- $\tau \in \left\{ 1, 2^{-1}, 10^{-1}, 10^{-2}, \right. \\ \left. 10^{-4}, 10^{-8}, 0 \right\}$
- 3 runs each; 95% CI

Avoiding inter- and intra-process comm.

- Do two reductions:
 - 1) Scale all but the diagonal element by τ
 - 2) Scale remote elements by τ (as before)

Avoiding inter- and intra-process comm.

- Do two reductions:
 - 1) Scale all but the diagonal element by τ
 - 2) Scale remote elements by τ (as before)
- If (1) gives the diagonal element, use it.
- Else use result of (2).

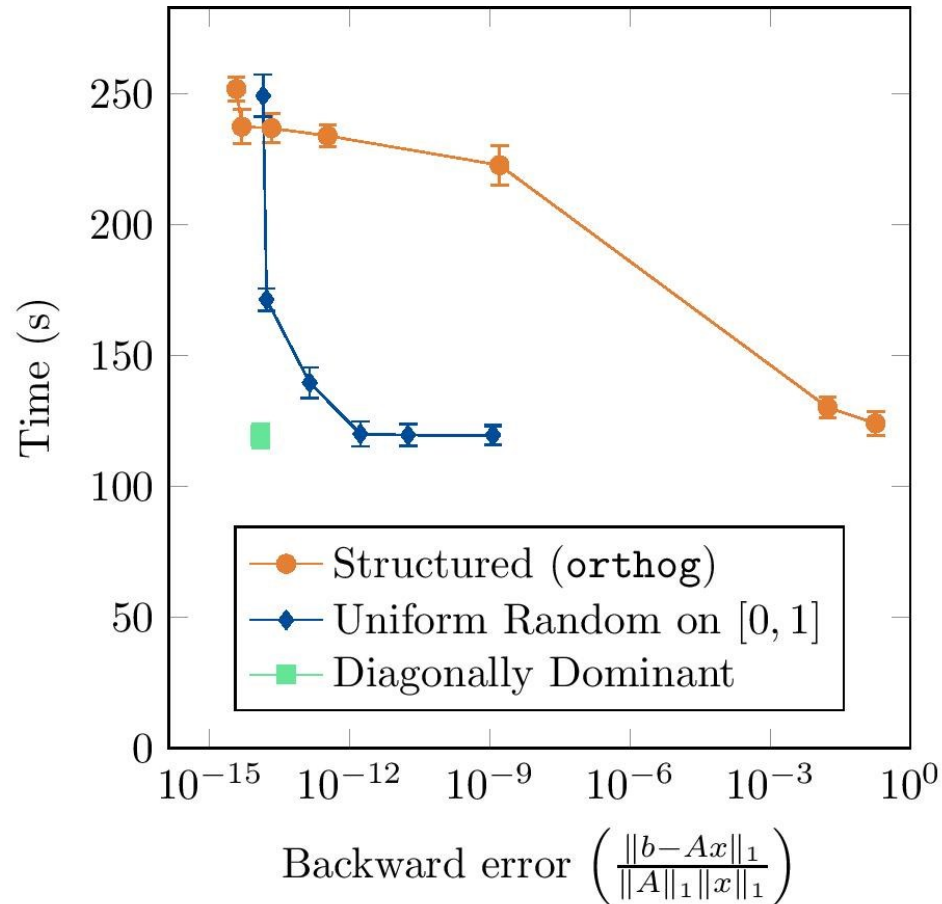
Avoiding inter- and intra-process comm.

- Do two reductions:
 - 1) Scale all but the diagonal element by τ
 - 2) Scale remote elements by τ (as before)
- If (1) gives the diagonal element, use it.
- Else use result of (2).

⇒ Selected pivot

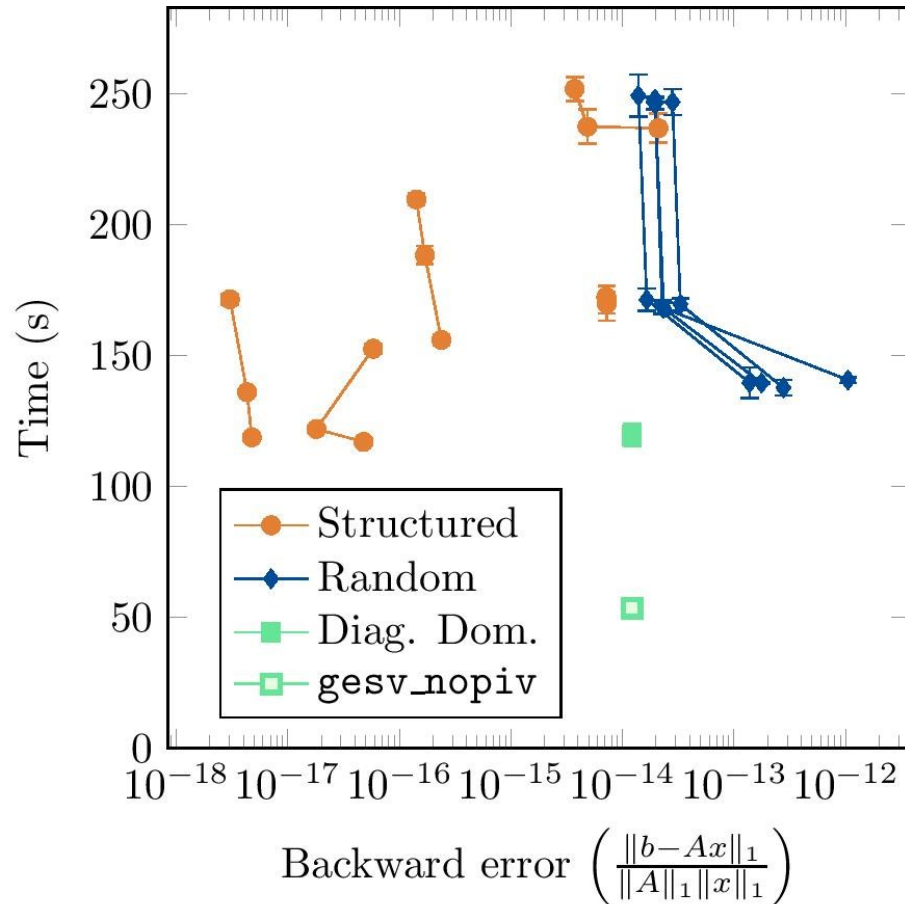
- Within τ of maximum
- Minimizes communication

Effect on performance



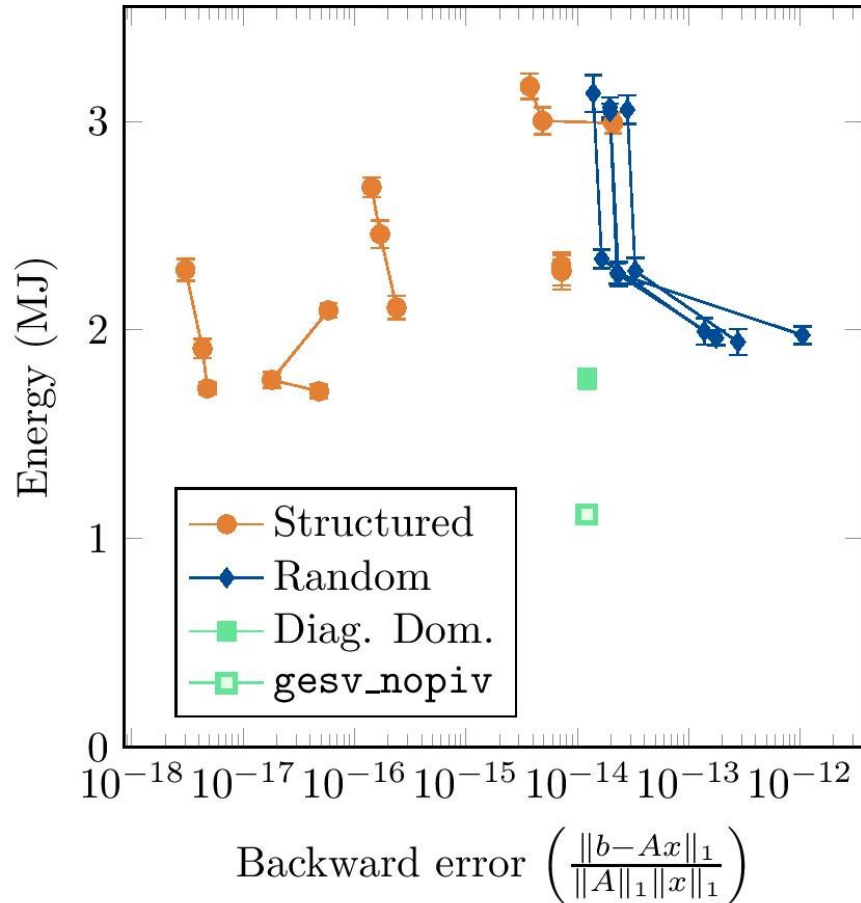
- 8 nodes of Summit
- SLATE w/ target=device
- $n = 225\,000$; nrhs = 10
- Double precision
- $\tau \in \left\{ 1, 2^{-1}, 10^{-1}, 10^{-2}, \right. \\ \left. 10^{-4}, 10^{-8}, 0 \right\}$
- 3 runs each; 95% CI

Effect on performance



- 8 nodes of Summit
- SLATE w/ target=device
- $n = 225\,000$; nrhs = 10
- Double precision
- $\tau \in \{1, 2^{-1}, 10^{-1}\}$
- 3 runs each; 95% CI

Effect on energy consumption



- 8 nodes of Summit
- SLATE w/ target=device
- $n = 225\,000$; nrhs = 10
- Double precision
- $\tau \in \{1, 2^{-1}, 10^{-1}\}$
- 3 runs each; 95% CI
- Energy measured w/ PAPI

Conclusions

- Threshold pivoting can reduce pivoting overhead
 - Without much loss of accuracy
 - Minor addition to partial pivoting
 - Already added to SLATE's LU
- ⇒ Valuable addition to distributed, dense LU code



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE